基于参与者共现分析的博文聚类研究*

龚凯乐　成　颖　孙建军
(南京大学信息管理学院　南京 210023)

摘要：【目的】将博文参与者共现作为特征，探析其在博文聚类中的价值。【方法】两步聚类：构建不同博文参与者的共现矩阵并转化为相关矩阵，采用近邻传播(Affinity Propagation, AP)算法完成第一步聚类；将 AP 聚类结果的质心作为初始聚类中心，对词项进行位置加权，利用 K-means 算法完成博文内容的第二步聚类。【结果】综合博文参与者共现与词项位置加权的聚类算法平均准确率与纯度分别达到 0.66 和 0.57，显著优于对比实验。【局限】本研究的主要贡献是引入参与者共现作为特征改进博文聚类效果，对于该特征甚少的博文聚类价值有限。【结论】整合词项与博文参与者特征的博文聚类显著地提高了聚类质量，两步法聚类也为 K-means 算法初始聚类中心的选择提供了可行的解决方案。

关键词：共现分析　文本聚类　博文参与者　初始聚类中心
分类号：G353

1　引　言

1973 年，Small[1]提出了共引理论，该理论的核心是共引能反映两篇文献在内容方面的相似性，共引关系的测度可用于揭示科学结构。其后，学者依据共现思想从作者、关键词等角度，进行了作者同被引[2]以及共词分析[3]。随着 Internet 的普及，Larson[4]将该思想推广到 Web 中开展了共链分析，用于揭示不同网站间的内容主题关系。目前，从共引、共词以及共链等具体应用中抽象出来的共现分析已经成为文献、信息以及科学计量学的重要研究方法，通过它可以发现研究对象间的亲疏关系，挖掘潜在知识，揭示对象所代表主体的结构与变化[5-8]。

在共现分析的应用方面，有学者将词汇[9-10]、引文[11]、链接[12]等特征的共现应用于文本聚类，改善了聚类质量。作者同被引分析能够划分作者群、确定领域核心作者[13]。社会化媒体中的共推荐关系[14]体现了用户共同兴趣、相似背景等同质性特征，为构建社区网络提供了依据。相关研究[15-17]发现，用户兴趣的稳定性使其在某段时间内较可能参与同一主题的交流，而通过长期互动、相同背景用户推荐等方式建立起的关注、好友等关系能让更多兴趣相投的用户加入到话题的讨论中。

作为社交媒体的用户，笔者发现博客中一个有趣的现象，即在相同或者相近主题的博文中，常会出现相当数量的共同参与者。共现思想以及文献[9-12]等的研究成果给笔者的启示是：博文参与者的共现应该可以作为博文聚类的特征。文献[15-17]的研究成果则提供了有益的实证研究基础，提示了博文参与者的共现与博文主题之间存在相关关系。据此，本文将参与者共现强度用于测度博文的主题相似度，通过共现分析挖掘出隐含的主题关联。

本文将参与博文撰写、评论、推荐的用户定义为该博文的参与者集合。采用两步法完成聚类：

(1) 构建不同博文的参与者共现矩阵并将其转化为相关矩阵，采用近邻传播 AP 算法[18]完成第一步聚类。

(2) 将 AP 聚类的质心作为初始聚类中心，利用 K-means 算法以词项位置加权为特征完成第二步聚类。

通讯作者：龚凯乐，ORCID：0000-0001-9269-8669，E-mail：njugong@163.com。

*本文系国家自然科学基金面上项目“融合范式视角下的链接分析理论集成框架及其实证研究”(项目编号：71273125)和中国科学技术信息研究所合作研究项目的研究成果之一。

实验证明，本研究中的博文参与者共现分析能有效地抽取出博文中隐含的主题关联，第一步聚类的质心有效地解决了 K-means 算法初始聚类中心不稳定的问题[19]。本文的研究工作为共现分析在信息检索、文本挖掘和知识发现的应用提供新的、有益的证据。

2 相关研究

2.1 基于共词分析的文本聚类研究

共词(Co-word)分析的基本思路是统计词项在同一篇文献中成对出现的频次，并基于该特征完成聚类分析。Liu 等[9]将文档集中高频共现的词项作为特征，实现向量空间降维。常鹏等[20]通过关联规则算法抽取出同一语境中的两个词作为共现词组合，以此为向量元素表示文档。Zhang 等[10]提出的 CoHC 算法根据词项共现关系发现文档集中的频繁 2 元组，扩展为 n 元组后去除冗余并排序，据此获得候选聚类标签并构建基类(共享同一短语的文档构成一个基类)，通过基类合并完成聚类。以 Zhang 等[10]的工作为基础，肖欣延等[21]将搜索引擎返回结果中与用户所用关键词频繁共现的词项构成重要词项集，将每个词项出现过的文档合并为对应的基类，依据文档重叠度进行基类合并，并利用 HowNet 的语义本体计算类间相似度优化聚类结果。李枫林等[22]依据词频、位置等特征提取文档关键词，以文档间的关键词共现频率构建共词矩阵，采用层次聚类法完成标签聚类，再依据标签组合实现文档聚类。

2.2 基于共引、共链分析的文本聚类研究

除了基于共词这一内容特征的文本聚类外，学者还将共引、共链等外部特征应用于文本聚类研究。算法思路主要有三种：直接将共引、共链特征应用于聚类研究[23-24]；将基于共引、共链特征的相似度与文本相似度融合完成聚类[12,25-26]；两步法聚类，即先利用共引、共链特征完成第一步聚类，然后再采用内容特征完成第二步聚类[11]。表 1 列举了每种思路的主要研究成果。

表 1 基于共引、共链分析的文本聚类研究

思路	作者	文章题名	研究详细思路
直接基于引文或链接的共现关系进行文本聚类	Wang Y, Kitsuregawa M [23]	Link Based Clustering of Web Search Results	将 Web 搜索结果的共链分析用于度量文本的相似度。
	Mukhopadhyay D, Sing S R [24]	An Algorithm for Automatic Web-page Clustering Using Link Structures	将链接分析的单位由单一网页拓展为主题文档集，即引用该主题文档集的链接等同于单一网页。基于此进行共链分析并完成文本聚类。
将基于引文、链接共现关系的相似度与文本内容相似度融合进行聚类	He X, Zha H, Ding C H Q, et al [12]	Web Document Clustering Using Hyperlink Structures	将 Web 文本的链接关系、共引模式和文本内容相融合，提出新的相似度计算方法。
	Modha D S, Spangler W S [25]	Clustering Hypertext with Applications to Web Searching	通过构造文本向量和链接关系向量，将文本相似度和共链相似度加以融合。
	顾钧，郑晓东，张连明[26]	结合引文信息的生物医学文本聚类研究	提出一种融合文本内容信息和引文信息的聚类算法，其中引文相似度综合考虑了引用、被引和共引关系。
基于引文与文本的两步法聚类	吴夙慧，成颖，郑彦宁，等[11]	基于学术文献同被引分析的 K-means 算法改进研究	算法构建了学术文献的同被引矩阵，基于此采用层次聚类获得 K-means 算法所需的 K 值以及初始质心。

2.3 博文聚类研究

现有博文聚类研究多集中在包含题名、关键词、标签(Tagging)、正文等的文本特征的应用。Brooks 等[27]采用 TF-IDF 方法从博文中提取出少量词项作为标签，以标签为特征进行聚类。何文静等[28]利用 Web2.0 特有的社会化标注系统，以用户添加的标签作为特征完成聚类。Zhang 等[29]利用博客网站、用户、博文、分类、标签等多种信息对社会化标注系统进行扩展，改善了单纯以标签为聚类特征的不足。Chen 等[30]参考社会化标注的概念提出一种基于博客搜索提问关键词的聚类算法，通过收集用户搜索博文时的提问词，利用形式概念分析抽取出可以表征博文内容的概念，以此聚类。Li 等[31]提出多特征融合的聚类方法，即将标题、内容和评论三部分作为博文特征，并赋予评论较高权重进行聚类。在链接特征的应用方面，Kopel 等[32]利用博文的内容关系、基于 XFN 链接的博主社交网络关系

以及基于 WordNet 的博文主题关系计算博文相似度实现聚类。博文链接特征在其他方面的应用主要包括社区发现、博主推荐、博文信息传播[33-36]等。

2.4 启　示

上述研究的结果显示：采用共词、共引以及共链特征的聚类算法均取得了更优的效果。从而验证了Small[1]的共引分析思想，即频繁成对出现在相同语境的特征是较为稳定的组合，可以表达潜在的主题信息。因此，以不同博文参与者的共现信息为特征开展探索性的聚类研究是值得尝试的。

3 研究设计

本文借鉴吴凤慧等[11]的研究思路，提出一种面向博文的两步聚类算法，如图 1 所示。算法首先构建博文参与者的共现矩阵，经 Jaccard 系数转化为相关矩阵后采用 AP 聚类算法[18]进行共现分析；第二步将 AP 聚类的质心作为初始聚类中心，利用 K-means 算法依据位置加权的词项特征完成博文聚类。

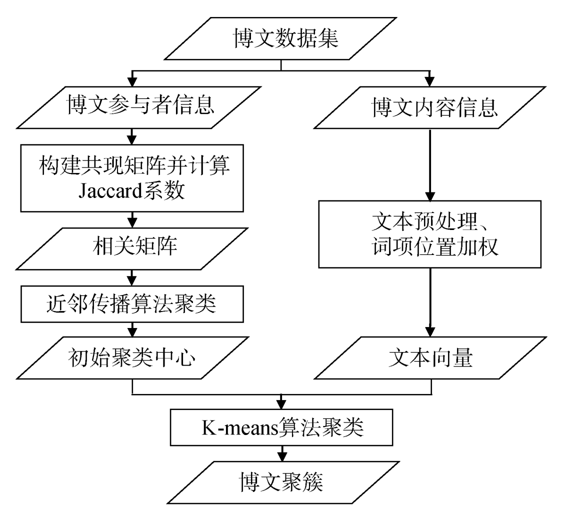


图 1　基于参与者共现分析改进的博文聚类流程

3.1 博文参与者共现分析

构建参与者共现矩阵如式(1)所示，其中 N_{ij} 表示同时出现在博文 i 和 j 中的参与者数量，当 i=j 时表示博文 i 的参与者数量，可简写为 N_i。

$$\begin{bmatrix} N_1 & N_{12} & \cdots & N_{1n} \\ N_{21} & N_2 & \cdots & N_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ N_{n1} & N_{n2} & \cdots & N_n \end{bmatrix} \tag{1}$$

利用 Jaccard 系数(公式(2))将共现矩阵转换为相关矩阵，如式(3)所示。J_{ij} 表示博文 i 和 j 基于参与者共现信息的相似度，J_{ij} 越大，博文 i 和 j 的相似度越高。

$$J_{ij} = N_{ij} \div (N_i + N_j - N_{ij}) \tag{2}$$

$$\begin{bmatrix} 1 & J_{12} & \cdots & J_{1n} \\ J_{21} & 1 & \cdots & J_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ J_{n1} & J_{n2} & \cdots & 1 \end{bmatrix} \tag{3}$$

(1) AP 算法

本文采用 AP 聚类算法[18]进行共现分析，它通过迭代更新样本空间中的相似关系信息确定每个样本的最优代表点，从而形成多个类簇。算法一般以 $n \times n$ 的相似度矩阵作为输入信息，并将对角线元素赋为同一值 p，称为偏向参数。p 值大小决定了每个样本最初被选作簇代表点的可能性，与簇的数量成正相关，因此可通过调整 p 值设定簇的数量[37]。算法通过迭代更新代表度 r(i, k)和适选度 a(i, k)实现聚类，采用 Jaccard 相似系数的 AP 算法流程如下：

①定义相似度矩阵元素 s(i, k)为公式(4)。令 a(i, k)=0。

$$s(i,k) = \begin{cases} J_{ik}, & i \neq j \\ p, & i = k \end{cases} \tag{4}$$

②依据公式(5)和公式(6)对 r(i, k)和 a(i, k)进行更新。

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t. k' \neq k}\{a(i,k') + s(i,k')\} \tag{5}$$

$$a(i,k) = \begin{cases} \min\{0, r(k,k) + \sum\limits_{i' s.t. i' \notin \{i,k\}} \max\{0, r(i',k)\}\}, & i \neq k \\ \sum\limits_{i' s.t. i' \neq k} \max\{0, r(i',k)\}, & i = k \end{cases} \tag{6}$$

③引入阻尼系数 λ，由公式(7)消除可能出现的震荡。

$$\begin{cases} r_{new}(i,k) = \lambda \times r_{old}(i,k) + (1-\lambda) \times r(i,k) \\ a_{new}(i,k) = \lambda \times a_{old}(i,k) + (1-\lambda) \times a(i,k) \end{cases} \tag{7}$$

其中，λ(0<λ<1)越大，其对震荡的消除效果越佳，一般将其设为 0.9[37]。

④通过公式(8)确定每一个样本 i 的簇代表点 c_i。

$$c_i \leftarrow \arg\max_{1 \leq k \leq n}\{a(i,k) + r(i,k)\} \tag{8}$$

⑤当迭代更新 10 次后所有样本的簇代表点仍保持不变或达到预先设置的最大迭代次数时算法结束，否则转到步骤②继续执行。

(2) 选取依据

周磊等[38]认为共现分析中经常采用的 K-means 以及层次聚类等算法存在缺陷，原因是它们均将“样本–样本”形式的相关矩阵理解为“样本–变量”形式的二维表，据此构建向量空间模型完成聚类。相关矩阵实质上已经体现了样本基于共现关系的相似度，若在聚类时以“样本–变量”形式计算向量相似度，划分类簇的

依据将不再是直接的共现关系，因此可以认为二者在理念上存在本质区别。实验证明基于 Pearson 系数采用欧氏距离进行层次聚类的共现分析效果不佳[38]。相较而言，本文采用的 AP 算法以相关矩阵为输入信息，以样本共现信息的相似关系为聚类依据，有效地改善了上述缺陷。同时，AP 算法将所有样本作为潜在中心，考虑了每个样本与全部样本的相似关系，不会收敛于局部最优。相关研究还表明，AP 算法比现有多数算法均优质、稳定[39]。其质心直观(聚类中心是确切存在的具体样本)的优点也符合本文 K-means 算法选择初始中心的要求。

3.2 基于词项位置加权的 K-means 算法

第二步聚类选用题名、标签、正文和评论作为特征构建向量空间模型，采用 K-means 算法完成聚类。由于上述 4 种内容特征表达主题的能力各异，因此赋予不同位置的特征项以不同权重。位置加权策略在学术界还没有统一规范，不同研究差别较大。韩客松等[40]认为学术文献标题中的特征项比摘要重要 3-5 倍，比正文重要 10-15 倍。Li 等[31]的研究表明加入评论能改进博客文本聚类质量，但苗家等[41]通过统计分析发现博客评论中存在大量噪声，且会出现主题漂移。郭朋伟等[42]通过实验发现博客题名、标签、正文和评论的权值分别取 2, 2, 3, 2 时，能取得较为满意的聚类结果。经观察与实验，本文将题名、标签、正文和评论的权重设为 2: 3: 2: 1，采用位置加权的 TF-IDF 方法对特征词项进行加权，公式如下：

$$w_{t,d} = TF_{t,d} \times IDF_t \tag{9}$$

其中，$TF_{t,d}$ 定义为词项 t 在文档 d 中的位置加权词频，IDF_t 称为逆文档频率。

$$TF_{t,d} = \lambda_t \times tf_{tt} + \lambda_k \times tf_{tk} + \lambda_c \times tf_{tc} + \lambda_r \times tf_{tr} \tag{10}$$

$$IDF_t = \log \frac{N}{DF_t} \tag{11}$$

公式(10)中，λ_t、λ_k、λ_c、λ_r 分别定义为题名、标签、正文和评论的权重系数，tf_{tt}、tf_{tk}、tf_{tc}、tf_{tr} 分别代表词项 t 在文档 d 的标题、标签、正文和评论中出现的次数。公式(11)中，N 指数据集中所有文档的数目，DF_t 称为文档频率，表示数据集中所有出现过词项 t 的文档的数目。

将 3.1 节确定的 K 个质心作为文本聚类 K-means 算法的初始中心，K-means 算法详见文献[43]。向量间的相似度计算方法采用欧氏距离，如下所示：

$$Sim(d_i, d_j) = \sqrt{\sum_{k=1}^{n} (W_{t_k,d_i} - W_{t_k,d_j})^2} \tag{12}$$

4 实验与结果分析

4.1 实验数据

科学网博客①是国内最为活跃的学术博客社区，笔者通过编写爬虫程序获取了截止 2015 年 12 月 19 日的 600 篇“半年内热门博文”内容及其参与者信息。在文本预处理中，调用 NLPIR 汉语分词系统②进行分词和词性标注。韩普等[44]的研究表明，采用名词、动词、形容词和副词的组合特征能比其他词性组合取得更好的聚类效果，因此过滤掉不属于该词性组合的文本内容。经过预处理后的博文和参与者信息共同构成了实验数据。两位研究人员分别独立对博文内容进行人工分类，并对首次分类不一致的博文征求第三人意见后归类，分类情况如表 2 所示：

表 2 学术博客人工分类表

编号	类名	数目	编号	类名	数目
1	学术科普	76	2	诺奖相关	45
3	评价考核	62	4	人物纪事	49
5	科研心得	140	6	教育教学	101
7	生活其他	102	8	时事资讯	25

4.2 评价指标

本文采用平均准确率[45]和纯度(Purity)[46]评价博文聚类质量。

平均准确率考察的是任意两篇博文人工分类和自动聚类结果的一致性。相关定义见表 3，其中积极准确率 PA 见公式(13)，消极准确率 NA 见公式(14)，平均准确率 AA 见公式(15)。显然，平均准确率越高，聚类质量越好。

①http://blog.sciencenet.cn/blog.php.
②http://ictclas.nlpir.org/.

表 3　分类聚类一致性关系及其数量符号

		手工分类中属于一个类	
		是	否
自动聚类中属于一个簇	是	积极正确数(TP)	积极错误数(FP)
	否	消极错误数(FN)	消极正确数(TN)

$$PA = \frac{TP}{TP + FN} \quad (13)$$

$$NA = \frac{TN}{FP + TN} \quad (14)$$

$$AA = \frac{PA + NA}{2} \quad (15)$$

纯度是另一种常用的聚类质量评价指标，计算时将各簇中的博文标注为该簇中数量最多的博文所在的类号，再通过正确标注的博文数除以博文集合中的文档总数求得聚类精度，公式如下：

$$Purity(\Omega, C) = \frac{1}{N}\sum_{k} \max_j |\omega_k \cap c_j| \quad (16)$$

其中，$\Omega = \{\omega_1, \omega_2, \cdots, \omega_k\}$ 为博文聚簇集合，$C = \{c_1, c_2, \cdots, c_j\}$ 是人工分类集合。

4.3　结果分析

本文的主要研究目的是探索不同博文特征对于聚类效果的影响，基于此，具体实验包括：

实验 1：参与者共现分析(AP 算法①)+题名、标签、正文和评论特征的词项位置加权(K-means 算法)；

实验 2：参与者共现分析(AP 算法)+题名、正文特征的 TF-IDF 加权算法(K-means 算法)；

实验 3：题名、标签、正文和评论特征的词项位置加权(AP 算法)；

实验 4：题名、标签、正文和评论特征的词项位置加权(K-means 算法)，其中经典 K-means 算法独立运行三次，聚类质量取均值。

(1) 不同文本特征对于聚类效果的影响

不同文本特征对于聚类效果的影响可以通过实验 1 和实验 2 的结果分析得到，如表 4 所示：

表 4　实验 1 与实验 2 的聚类结果

实验	积极准确率 PA	消极准确率 NA	平均准确率 AA	纯度 Purity
实验 1	0.460940	0.866579	0.663760	0.571667
实验 2	0.397634	0.776111	0.586873	0.413333

从表 4 中的 4 个评测指标均可以看出，在聚类算法相同的情况下，采用博文词项位置加权比单纯以题名和正文为特征的 TF-IDF 加权方法具有更优的质量。这也证明了在对博客等社交媒体中的文本进行聚类时，引入标签、评论等内容并赋予其相应的特征权重具有明显的价值。其原因也不难理解，标签是作者对博文内容的凝练概括，类似学术文献中的关键词，对表达文本主题、区分文本内容具有明确的意义。而评论多是读者与作者就博文内容展开的互动与讨论，存在大量主题相关的特征词项，特别是在以图片、音乐、视频等多媒体形式存在的博客中，评论对文本聚类的指导意义显得更加重要。

(2) 参与者共现信息对于博文聚类的价值

参与者共现信息对于博文聚类的价值可以通过实验 1、实验 3 以及实验 4 的分析比较得出，结果如表 5 所示：

表 5　实验 1、实验 3 与实验 4 的聚类结果

实验		积极准确率 PA	消极准确率 NA	平均准确率 AA	纯度 Purity
实验 1		0.460940	0.866579	0.663760	0.571667
实验 3		0.316383	0.854485	0.585434	0.496667
实验 4	1	0.382600	0.782003	0.582302	0.400000
	2	0.380831	0.769077	0.574954	0.446667
	3	0.395055	0.751616	0.573335	0.450000
	平均	0.386162	0.767565	0.576864	0.432222

通过表 5 可以得到以下结论：

①参与者共现特征明显改进聚类效果。表 5 中的 PA、NA、AA 以及纯度的结果均显示，在文本特征项和加权方法相同的情况下，本文提出的基于参与者共现分析的文本聚类算法明显优于基于文本内容的近邻传播聚类算法和经典 K-means 算法，证实了通过共现分析，能有效地将博文参与者共现信息中隐含的主题关联抽取出来。

②两步聚类算法改善了 K-means 算法初始聚类中心的确定。第一步聚类得到了优化的初始聚类中心，改进了聚类质量。

③AP 算法稳定质优。通过实验 3 与实验 4 的对比可以看出经典 K-means 算法独立运行三次的结果很不稳定，其随机选择初始聚类中心的一般做法对于聚类结果存在显著的影响，该做法使得聚类结果具有较大的偶然性，而 AP 算法

①http://www.psi.toronto.edu/index.php?q=affinity%20propagation.

对同一数据集的多次运行保持了稳定的聚类结果，该特点也保证了本研究所提算法的稳定性。从 AA 和纯度的数据还可以发现，AP 算法的准确性优于经典 K-means 算法。进一步观察 PA 和 NA，AP 算法的 PA 值虽然稍低，但 NA 却明显高于经典 K-means 算法。从而说明在本文这种类间区分度小的数据集上，AP 算法更能将不同类别的对象区分开来，形成较为均匀的多个类簇。

5 结 语

本文通过对参与者的共现分析改进了博文聚类效果。实验结果表明参与者共现信息中隐含的主题关联信息能够改进经典 K-means 算法初始聚类中心的选择，有效地提升了聚类效果以及稳定性；此外，博文词项位置加权以及文本特征的综合应用也有助于提升聚类质量。

目前，博文聚类研究是学界的研究热点，本文的工作为该领域提供了有益的思路。不过，本研究主要依据参与者共现特征改善了聚类效果，对于参与者较少的博文将导致相关矩阵变得十分稀疏，不利于第一步聚类。因此探索参与者稀疏矩阵背景下的博文聚类成为后续的研究主题。

参考文献：

[1] Small H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents [J]. Journal of the American Society for Information Science, 1973, 24(4): 265-269.

[2] White H D, Griffith B C. Author Cocitation: A Literature Measure of Intellectual Structure[J]. Journal of the American Society for Information Science, 1981, 32(3): 163-171.

[3] Callon M, Law J, Rip A. Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World[M]. Macmillan Press Ltd., 1986.

[4] Larson R R. Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace [C]. In: Proceedings of the Annual Meeting-American Society for Information Science.1996, 33: 71-78.

[5] 王曰芬，宋爽，卢宁，等. 共现分析在文本知识挖掘中的应用研究[J]. 中国图书馆学报，2007，33(2): 59-64. (Wang Yuefen, Song Shuang, Lu Ning, et al. Applications of Co-occurrence Analysis in Text Knowledge Mining [J]. Journal of Library Science in China, 2007, 33(2): 59-64.)

[6] 张树良，冷伏海. 基于文献的知识发现的应用进展研究[J]. 情报学报，2006，25(6): 700-712. (Zhang Shuliang, Leng Fuhai. Study on the Applicational Development of Literature-based Knowledge Discovery[J]. Journal of the China Society for Scientific and Technical Information, 2006, 25(6): 700-712.)

[7] 王曰芬，宋爽，苗露. 共现分析在知识服务中的应用研究[J]. 现代图书情报技术，2006(4): 29-34. (Wang Yuefen, Song Shuang, Miao Lu. Application Study of Co-occurrence Analysis in Knowledge Service [J]. New Technology of Library and Information Service, 2006(4): 29-34.)

[8] 孙建军，李江. 网络信息计量理论、工具与应用[M]. 北京: 科学出版社，2009.(Sun Jianjun, Li Jiang. On Webometrics Theories, Tools and Applications[M]. Beijing: Science Press, 2009.)

[9] Liu Y C, Wang X L, Liu B Q. A Feature Selection Algorithm for Document Clustering Based on Word Co-occurrence Frequency [C]. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics. IEEE, 2004, 5: 2963-2968.

[10] Zhang Y, Feng B Q. A Co-occurrence Based Hierarchical Method for Clustering Web Search Results[C]. In: Proceedings of IEEE / WIC / ACM International Conference on Web Intelligence, WI 2008, Sydney, Australia. 2008: 407-410.

[11] 吴夙慧，成颖，郑彦宁，等. 基于学术文献同被引分析的 K-means 算法改进研究[J]. 情报学报，2012，31(1): 82-94. (Wu Suhui, Cheng Ying, Zheng Yanning, et al. Improvement of K-means Algorithm Based on Co-citation Analysis[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(1): 82-94.)

[12] He X, Zha H, Ding C H Q, et al. Web Document Clustering Using Hyperlink Structures[J]. Computational Statistics & Data Analysis, 2002, 41(1): 19-45.

[13] 邓三鸿，顾婷婷. 我国图情领域核心期刊论文作者同被引现象的可视化分析[J]. 情报科学，2010，28(11): 1728-1732. (Deng Sanhong, Gu Tingting. A Visual ACA Analysis of Core Journals in the Field of LIS[J]. Information Science, 2010, 28(11): 1728-1732.)

[14] 谭旻，许鑫，赵星. 学术博客共推荐关系及核心结构特性研究——以科学网博客为例[J]. 现代图书情报技术，2015(7): 24-30. (Tan Min, Xu Xin, Zhao Xing. Exploring the Co-recommendation Relationship and Its Core Structure Features of Academic Blogs——Taking ScienceNet.cn Blog as an Example [J]. New Technology of Library and Information Service, 2015(7): 24-30.)

[15] Xia F, Yang Q, Li J, et al. Data Dissemination Using Interest-tree in Socially Aware Networking [J]. Computer Networks, 2015, 91: 495-507.

[16] McPherson M, Smith-Lovin L, Cook J M. Birds of a Feather: Homophily in Social Networks [J]. Annual Review of Sociology, 2001, 27(1): 415-444.

[17] Katsaros D, Dimokas N, Tassiulas L. Social Network Analysis Concepts in the Design of Wireless Ad Hoc Network Protocols[J]. IEEE Network, 2010, 24(6): 23-29.

[18] Frey B J, Dueck D. Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315(5814): 972-976.

[19] 吴夙慧，成颖，郑彦宁，等. K-means 算法研究综述[J]. 现代图书情报技术，2011(5): 28-35. (Wu Suhui, Cheng Ying, Zheng Yanning, et al. Survey on K-means Algorithm[J]. New Technology of Library and Information Service, 2011(5): 28-35.)

[20] 常鹏，冯楠，马辉. 一种基于词共现的文档聚类算法[J]. 计算机工程，2012, 38(2): 213-214. (Chang Peng, Feng Nan, Ma Hui. Document Clustering Algorithm Based on Word Co-occurrence [J]. Computer Engineering, 2012, 38(2): 213-214.)

[21] 肖欣延，张东站，高君杰，等. 一种新的 Web 检索结果聚类方法[J]. 计算机研究与发展，2007, 44(S2): 79-83. (Xiao Xinyan, Zhang Dongzhan, Gao Junjie, et al. A New Method for Web Search Results Clustering [J]. Journal of Computer Research and Development, 2007, 44(S2): 79-83.)

[22] 李枫林，何洲芳. 基于关键词共现分析的检索结果聚类研究[J]. 情报学报，2011, 30(8): 819-825. (Li Fenglin, He Zhoufang. Study on Clustering of Retrieval Results Based on Co-occurrence Analysis of Keywords [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(8): 819-825.)

[23] Wang Y, Kitsuregawa M. Link Based Clustering of Web Search Results [C]. In: Proceedings of International Conference on Advances in Web-Age Information Management. Springer-Verlag, 2001: 225-236.

[24] Mukhopadhyay D, Sing S R. An Algorithm for Automatic Web-page Clustering Using Link Structures [C]. In: Proceedings of the IEEE INDICON Annual Conference 2004. IEEE, 2004: 472-477.

[25] Modha D S, Spangler W S. Clustering Hypertext with Applications to Web Searching [C]. In: Proceedings of the 11th ACM Conference on Hypertext and Hypermedia. ACM, 2000: 143-152.

[26] 顾钧，郑晓东，张连明. 结合引文信息的生物医学文本聚类研究[J]. 计算机应用与软件，2012(10): 5-7. (Gu Jun, Zheng Xiaodong, Zhang Lianming. Research on Bio-medical Document Clustering with Citation Information Incorporated [J]. Computer Applications and Software, 2012(10): 5-7.)

[27] Brooks C H, Montanez N. Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering [C]. In: Proceedings of the 15th International Conference on World Wide Web. ACM, 2006: 625-632.

[28] 何文静，何琳. 基于社会标签的文本聚类研究[J]. 现代图书情报技术，2013(7-8): 49-54. (He Wenjing, He Lin. Research on Text Clustering Based on Social Tagging [J]. New Technology of Library and Information Service, 2013(7-8): 49-54.)

[29] Zhang Y, Gao K, Zhang B, et al. Clustering Blog Posts Using Tags and Relations in the Blogosphere [C]. In: Proceedings of the 1st International Conference on Information Science and Engineering (ICISE). IEEE, 2010: 817-820.

[30] Chen Y H, Lu J L, Wu T Y. A Blog Clustering Approach Based on Queried Keywords [C]. In: Proceedings of the 2013 International Symposium on Biometrics and Security Technologies (ISBAST). IEEE, 2013: 1-9.

[31] Li B, Xu S, Zhang J. Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments [C]. In: Proceedings of the 45th Annual Southeast Regional Conference. ACM, 2007: 94-99.

[32] Kopel M, Zgrzywa A. Search Result Clustering Using Semantic Web Data [C]. In: Proceedings of the 3rd International Conference on Intelligent Information and Database Systems. Springer Berlin Heidelberg, 2011: 292-301.

[33] Chin A, Chignell M. A Social Hypertext Model for Finding Community in Blogs [C]. In: Proceedings of the 17th Conference on Hypertext and Hypermedia. ACM, 2006: 11-22.

[34] Lin Y R, Sundaram H, Chi Y, et al. Discovery of Blog Communities Based on Mutual Awareness [C]. In: Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem. 2006.

[35] Lu L, Zhu F. Blogger Clustering by Utilizing Link Information [C]. In: Proceedings of the 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS). IEEE, 2010, 2: 267-270.

[36] Bruns A, Burgess J, Highfield T, et al. Mapping the Australian Networked Public Sphere [J]. Social Science Computer Review, 2011, 29(3): 277-287.

[37] 肖宇，于剑. 基于近邻传播算法的半监督聚类[J]. 软件学

报, 2008, 19(11): 2803-2813. (Xiao Yu, Yu Jian. Semi-Supervised Clustering Based on Affinity Propagation Algorithm[J]. Journal of Software, 2008, 19(11): 2803-2813.)

[38] 周磊, 杨威, 张玉峰. 共现矩阵聚类分析的问题与再思考[J]. 情报杂志, 2014, 33(6): 32-36. (Zhou Lei, Yang Wei, Zhang Yufeng. Issues and Re-consideration on Cluster Analysis in Co-occurrence Matrix[J]. Journal of Intelligence, 2014, 33(6): 32-36.)

[39] 杭文龙, 蒋亦樟, 刘解放, 等. 迁移近邻传播聚类算法[J/OL]. 软件学报, (2015-11-26).[2016-04-01]. http://www.cnki.net/kcms/detail/11.2560.TP.20151126.1606.001.html. (Hang Wenlong, Jiang Yizhang, Liu Jiefang, et al. Transfer Affinity Propagation Clustering Algorithm[J/OL]. Journal of Software, (2015-11-26). [2016-04-01]. http://www.cnki.net/kcms/detail/11.2560.TP.20151126.1606.001.html.)

[40] 韩客松, 王永成. 中文全文标引的主题词标引和主题概念标引方法[J]. 情报学报, 2001, 20(2): 212-216. (Han Kesong, Wang Yongcheng. Methods of Keyword and Subject Concept Indexing to Chinese Full-text[J]. Journal of the China Society for Scientific and Technical Information, 2001, 20(2): 212-216.)

[41] 苗家, 马军, 陈竹敏. 一种基于HITS算法的Blog文摘方法[J]. 中文信息学报, 2011, 25(1): 104-109. (Miao Jia, Ma Jun, Chen Zhumin. A New HITS-Based Summarization Approach for Blog[J]. Journal of Chinese Information Processing, 2011, 25(1): 104-109.)

[42] 郭朋伟, 高克宁, 张斌. 基于评论修正的博客聚类算法[J]. 东北大学学报: 自然科学版, 2010, 31(6): 782-785. (Guo Pengwei, Gao Kening, Zhang Bin. Public Blog Clustering Algorithm Based on Revision by Comments[J]. Journal of Northeastern University: Natural Science, 2010, 31(6): 782-785.)

[43] MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967, 1: 281-297.

[44] 韩普, 王东波, 刘艳云, 等. 词性对中英文文本聚类的影响研究[J]. 中文信息学报, 2013, 27(2): 65-73. (Han Pu, Wang Dongbo, Liu Yanyun, et al. Influence of Part-of-Speech on Chinese and English Document Clustering[J]. Journal of Chinese Information Processing, 2013, 27(2): 65-73.)

[45] 王娟, 范少萍, 郑春厚. 基于惩罚性矩阵分解的文本聚类分析[J]. 情报学报, 2012, 31(9): 998-1008. (Wang Juan, Fan Shaoping, Zheng Chunhou. Penalized Matrix Decomposition Method for Text Clustering[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(9): 998-1008.)

[46] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval[M]. Cambridge: Cambridge University Press, 2008.

作者贡献声明:

龚凯乐: 设计研究方案, 采集数据, 完成实验, 撰写论文初稿;
成颖: 提出研究思路, 优化研究方案, 论文修订;
孙建军: 提出论文撰写框架。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 龚凯乐, 成颖, 孙建军. Participants.zip. 博文参与者数据.

[2] 龚凯乐, 成颖, 孙建军. ParticipantCorrelationMatrix.txt. 博文参与者相关矩阵.

[3] 龚凯乐, 成颖, 孙建军. BlogText.zip. 预处理后的博文内容数据.

[4] 龚凯乐, 成颖, 孙建军. PositionWeightedVSM.txt. 题名、标签、正文和评论词项位置加权的向量空间矩阵.

[5] 龚凯乐, 成颖, 孙建军. TitleBodyTFIDFVSM.txt. 题名、正文TF-IDF 加权的向量空间矩阵.

[6] 龚凯乐, 成颖, 孙建军. ExperimentalData.xlsx. 人工分类情况与聚类结果.

收稿日期: 2016-05-04
收修改稿日期: 2016-06-13

Clustering Blog Posts with Co-occurrence Analysis

Gong Kaile　Cheng Ying　Sun Jianjun
(School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract: [**Objective**] This study investigates the co-occurrence of blog comment contributors, aiming to explore their roles in blog posts clustering. [**Methods**] We developed a method of two-step clustering. First, we constructed the co-occurrence matrix of the contributors from different blog posts and then transform it to a correlation matrix. Then finished the first-step clustering with the help of Affinity Propagation (AP) algorithm. Second, we calculated the terms' position weight based on the centers of AP clustering, and then finished the second-stage blog post content clustering with K-means algorithm. [**Results**] The average precision and recall ratio of the proposed method were 0.66 and 0.57, which were significantly higher than those of the traditional ones. [**Limitations**] The blog comment contributors co-occurrence improved the quality of clustering, but it has limited value in blog posts with few comments. [**Conclusions**] The proposed method improves the quality of blog posts clustering by combining terms and contributors' co-occurrence. The two-step clustering method is a better option to select the initial cluster centers of the K-means algorithm.

Keywords: Co-occurrence analysis　Text clustering　Blog comments contributor　Initial cluster centers